

On the Confidentiality of Information Dispersal Algorithms and Their Erasure Codes

Mingqiang Li

Department of Computer Science and Engineering, The Chinese University of Hong Kong
Shatin, New Territories, Hong Kong
Email: mingqiangli.cn@gmail.com

Abstract—*Information Dispersal Algorithms (IDAs)* have been widely applied to reliable and secure storage and transmission of data files in distributed systems. An IDA is a method that encodes a file F of size $L = |F|$ into n unrecognizable pieces F_1, F_2, \dots, F_n , each of size L/m ($m < n$), so that the original file F can be reconstructed from any m pieces. The core of an IDA is the adopted non-systematic m -of- n erasure code. This paper makes a systematic study on the *confidentiality* of an IDA and its connection with the adopted erasure code. Two levels of confidentiality are defined: *weak confidentiality* (in the case where some parts of the original file F can be reconstructed explicitly from fewer than m pieces) and *strong confidentiality* (in the case where nothing of the original file F can be reconstructed explicitly from fewer than m pieces). For an IDA that adopts an arbitrary non-systematic erasure code, its confidentiality may fall into weak confidentiality. To achieve strong confidentiality, this paper explores a sufficient and feasible condition on the adopted erasure code. Then, this paper shows that Rabin's IDA has strong confidentiality. At the same time, this paper presents an effective way to construct an IDA with strong confidentiality from an arbitrary m -of- $(m+n)$ erasure code. Then, as an example, this paper constructs an IDA with strong confidentiality from a Reed-Solomon code, the computation complexity of which is comparable to or sometimes even lower than that of Rabin's IDA.

Index Terms—Cauchy matrix, confidentiality, erasure code, information dispersal algorithm, Reed-Solomon code, Vandermonde matrix.

I. INTRODUCTION

In 1989, Rabin [1] proposed an attractive *Information Dispersal Algorithm (IDA)* that is applicable to reliable and secure storage and transmission of data files in distributed systems. Since then, IDAs have drawn many attentions from both researchers and engineers in the area of distributed systems.

An IDA is a method that encodes a file F of size $L = |F|$ into n unrecognizable pieces F_1, F_2, \dots, F_n , each of size L/m ($m < n$), so that the original file F can be reconstructed from any m pieces. From a coding theorist's viewpoint, an IDA is corresponding to a non-systematic m -of- n erasure code [2]. Here, the non-systematic property of the erasure code is necessary to ensure “unrecognizable” pieces. In practice, an IDA is implemented as follows: The original file F is firstly divided into m segments S_1, S_2, \dots, S_m , each of size L/m . Then, the m segments are encoded into n unrecognizable pieces F_1, F_2, \dots, F_n using a non-systematic m -of- n erasure

code.

The reliability of an IDA is clear: no more than $n - m$ lost pieces of the n pieces F_1, F_2, \dots, F_n will not result in data loss. However, the *confidentiality* of an IDA is not straightforward and deserves a systematic study.

From the view of information-theoretic security [3], IDAs can provide only incremental confidentiality and thus have weaker confidentiality than secret sharing (with perfect confidentiality) [4]–[6] and ramp schemes (with partially perfect confidentiality) [7], [8]. However, IDAs can achieve optimal efficiency in data overhead [1]. As shown in [9, Page 65, Table A], there is a trade-off between confidentiality and data overhead. Moreover, for practical applications, information-theoretic security is often extravagant and unnecessary. Thus, in this paper, we will study the practical security that IDAs can provide.

Although a non-systematic erasure code can ensure “unrecognizable” pieces in an IDA, some segments may still be reconstructed explicitly from fewer than m pieces. Then, an eavesdropper who acquires fewer than m pieces by snooping may reconstruct some parts of the original file F explicitly, resulting in partial file leakage. In the case of partial file leakage, we say the IDA has *weak confidentiality*. However, for an ideal IDA, any segment of the original file F should not be reconstructed explicitly from fewer than m pieces. In the case where nothing of the original file F can be reconstructed explicitly from fewer than m pieces, we say the IDA has *strong confidentiality*.

For an IDA that adopts an arbitrary non-systematic erasure code, we noticed that its confidentiality may fall into weak confidentiality. In this paper, we will first show in Section III which kind of IDAs has weak confidentiality and how an eavesdropper can reconstruct some segments of the original file F explicitly from fewer than m pieces in the case of weak confidentiality. Then, to achieve strong confidentiality, we explore a sufficient and feasible condition for an IDA in Section IV. We show that Rabin's IDA [1] has strong confidentiality. At the same time, we present an effective way to construct an IDA with strong confidentiality from an arbitrary m -of- $(m+n)$ erasure code. Then, as an example, we construct an IDA with strong confidentiality from a Reed-Solomon code [10], the computation complexity of which is comparable to or sometimes even lower than that of Rabin's IDA. Finally, we conclude this paper in Section V. To our knowledge, this paper is the first work that focuses on the

The main part of this work was finished while Mingqiang Li worked as a Staff Researcher in the IBM China Research Laboratory.

issues of weak confidentiality and strong confidentiality in IDAs.

To make our later discussion more easily understood, we begin this paper with a brief introduction of IDAs and their erasure codes.

II. IDAS AND THEIR ERASURE CODES

In an Information Dispersal Algorithm (IDA), a non-systematic m -of- n erasure code is employed to encode the m segments S_1, S_2, \dots, S_m into n unrecognizable pieces F_1, F_2, \dots, F_n , i.e.

$$(S_1, S_2, \dots, S_m) \cdot G_{m \times n} = (F_1, F_2, \dots, F_n), \quad (1)$$

where $G_{m \times n}$ is the *generator matrix* of the adopted erasure code and meets the following two conditions:

- 1) Any column of $G_{m \times n}$ is not equal to any column of an $m \times m$ identity matrix; and
- 2) Any m columns of $G_{m \times n}$ form an $m \times m$ nonsingular matrix.

The first condition ensures that any piece is unrecognizable; while the second condition ensures that the original file F can be reconstructed from any m pieces.

III. WHICH KIND OF IDAS HAS WEAK CONFIDENTIALITY

In this section, we will show which kind of Information Dispersal Algorithms (IDAs) have weak confidentiality and how an eavesdropper can reconstruct some segments of the original file F explicitly from fewer than m pieces in the case of weak confidentiality. We present a theorem as follows:

Theorem 3.1: An IDA has *weak confidentiality* if and only if the adopted erasure code meets the following condition: In its generator matrix $G_{m \times n}$, there is a submatrix $A_{m' \times n'}$ of column rank r , where $m', n' < m$ and $n' - r = m - m' > 0$.

Proof: We first prove the sufficiency. Suppose $A_{m' \times n'}$ is located in rows $i_1, i_2, \dots, i_{m'}$ and columns $j_1, j_2, \dots, j_{n'}$ of $G_{m \times n}$. Then, $S_{i_1}, S_{i_2}, \dots, S_{i_{m'}}$ are the m' segments corresponding to rows $i_1, i_2, \dots, i_{m'}$ of $G_{m \times n}$. Similarly, $F_{j_1}, F_{j_2}, \dots, F_{j_{n'}}$ are the n' pieces corresponding to columns $j_1, j_2, \dots, j_{n'}$ of $G_{m \times n}$. Since $n' - r = m - m' > 0$, $r = m' + n' - m$. Furthermore, since $m', n' < m$, then $r < m', n'$. Thus, $A_{m' \times n'}$ is rank deficient. Then, in $A_{m' \times n'}$, any $k = n' - r$ columns can be linearly represented by other r columns. Let $A_{m' \times n'} = (v_1, v_2, \dots, v_{n'})$, where $v_1, v_2, \dots, v_{n'}$ are column vectors. Suppose there is a linear relation among column vectors of $A_{m' \times n'}$ as follows:

$$(v_1, v_2, \dots, v_k) = (v_{k+1}, v_{k+2}, \dots, v_{n'}) \cdot B_{r \times k},$$

where $B_{r \times k}$ is the transpose of coefficient matrix. Then, any information of $S_{i_1}, S_{i_2}, \dots, S_{i_{m'}}$ can be eliminated by calculating

$$(\tilde{F}_{j_1}, \tilde{F}_{j_2}, \dots, \tilde{F}_{j_k}) = (F_{j_1}, F_{j_2}, \dots, F_{j_k}) - (F_{j_{k+1}}, F_{j_{k+2}}, \dots, F_{j_{n'}}) \cdot B_{r \times k}. \quad (2)$$

Finally, according to Equation (1), other $m - m' = k$ segments except $S_{i_1}, S_{i_2}, \dots, S_{i_{m'}}$ can be reconstructed explicitly.

We now prove the necessity. If some segments can be

reconstructed explicitly from fewer than m pieces in an IDA, it is clear that the information of other segments should be able to be eliminated from the eavesdropped pieces by linear operations. Moreover, a solvable system of linear equations on these reconstructible segments should be able to be formed. From the above proof of sufficiency, we can deduce that in the corresponding generator matrix $G_{m \times n}$, there should be a submatrix $A_{m' \times n'}$ of column rank r , where $m', n' < m$ and $n' - r = m - m' > 0$. ■

Remark: From the second condition in the previous section, we can deduce a necessary condition $n' - r \leq m - m'$ (otherwise the corresponding n' columns of $G_{m \times n}$ will form a rank deficient matrix—a contradiction). Thus, for an IDA that adopts an arbitrary non-systematic erasure code, its confidentiality may fall into weak confidentiality.

IV. CONSTRUCTING IDAS WITH STRONG CONFIDENTIALITY

For an Information Dispersal Algorithm (IDA), to achieve strong confidentiality, we explore a sufficient and feasible condition as follows:

Theorem 4.1: An IDA has strong confidentiality if the adopted erasure code meets the following condition: Any square submatrix of its generator matrix $G_{m \times n}$ is nonsingular.

Proof: We prove this theorem by contradiction as follows: In this case, suppose this IDA has weak confidentiality. According to Theorem 3.1, in $G_{m \times n}$, there is a submatrix $A_{m' \times n'}$ of column rank r , where $m', n' < m$ and $n' - r = m - m' > 0$. Then, according to the proof of Theorem 3.1, $A_{m' \times n'}$ is rank deficient. Thus, in $A_{m' \times n'}$, any $\min(m', n') \times \min(m', n')$ square submatrix is singular—a contradiction! Therefore, this IDA has strong confidentiality. ■

In Rabin's IDA [1], the corresponding generator matrix is a Cauchy matrix, in which any square submatrix is nonsingular. Thus, Rabin's IDA has strong confidentiality.

Inspired by the work in [11], we now present an effective way to construct an IDA with strong confidentiality from an arbitrary m -of- $(m+n)$ erasure code as follows:

- 1) Choose an arbitrary m -of- $(m+n)$ erasure code, whose generator matrix is $G_{m \times (m+n)} = (C_{m \times m} | D_{m \times n})$;
- 2) Construct an IDA that adopts an m -of- n erasure code whose generator matrix is $C_{m \times m}^{-1} \cdot D_{m \times n}$.

It is easy to verify that the IDA constructed above has strong confidentiality as follows:

- 1) In the case where $C_{m \times m}$ is an $m \times m$ identity matrix, the chosen m -of- $(m+n)$ erasure code is a systematic erasure code. Then, according to the nature of a systematic m -of- $(m+n)$ erasure code [2], any square submatrix of $D_{m \times n} = C_{m \times m}^{-1} \cdot D_{m \times n}$ is nonsingular. Thus, according to Theorem 4.1, the constructed IDA has strong confidentiality.
- 2) In the case where $C_{m \times m}$ is not an $m \times m$ identity matrix, the chosen m -of- $(m+n)$ erasure code is a non-systematic erasure code. Then, $(I_{m \times m} | C_{m \times m}^{-1} \cdot D_{m \times n})$ is the generator matrix of the equivalent systematic m -of- $(m+n)$ erasure

code. So, any square submatrix of $C_{m \times m}^{-1} \cdot D_{m \times n}$ is nonsingular. Thus, according to Theorem 4.1, the constructed IDA also has strong confidentiality.

Example: We construct an IDA with strong confidentiality from a Reed-Solomon code [10], whose generator matrix is a Vandermonde matrix. From what we have discussed above, we first choose a m -of- $(m+n)$ Reed-Solomon code with generator matrix

$$G_{RS} = \begin{pmatrix} a_1^0 & a_2^0 & \cdots & a_{m+n}^0 \\ a_1^1 & a_2^1 & \cdots & a_{m+n}^1 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{m-1} & a_2^{m-1} & \cdots & a_{m+n}^{m-1} \end{pmatrix}, \quad (3)$$

where a_1, a_2, \dots, a_{m+n} are distinct. Then, an IDA with strong confidentiality can be reconstructed, in which the corresponding generator matrix is

$$G_{IDA} = \begin{pmatrix} a_1^0 & a_2^0 & \cdots & a_m^0 \\ a_1^1 & a_2^1 & \cdots & a_m^1 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{m-1} & a_2^{m-1} & \cdots & a_m^{m-1} \end{pmatrix}^{-1} \times \begin{pmatrix} a_{m+1}^0 & a_{m+2}^0 & \cdots & a_{m+n}^0 \\ a_{m+1}^1 & a_{m+2}^1 & \cdots & a_{m+n}^1 \\ \vdots & \vdots & \ddots & \vdots \\ a_{m+1}^{m-1} & a_{m+2}^{m-1} & \cdots & a_{m+n}^{m-1} \end{pmatrix}. \quad (4)$$

From the comparison results in [12], we can deduce that the computation complexity of this IDA is comparable to or sometimes even lower than that of Rabin's IDA.

Remark: Besides Cauchy matrices, Vandermonde matrices were also suggested for the generator matrices of IDAs in Rabin's seminal paper [1, Page 339]. However, a Vandermonde matrix defined over a finite field may contain singular square submatrices [2, Page 323, Problem. (7)]. Then, an IDA whose erasure code is a Reed-Solomon code defined over a finite field may not meet the condition in Theorem 4.1 and thus may have weak confidentiality. Luckily, in the literature, when Rabin's IDA is mentioned, it always refers to that constructed based on a Cauchy matrix.

V. CONCLUSIONS

Information Dispersal Algorithms (IDAs) [1] have been widely applied to reliable and secure storage and transmission of data files in distributed systems. This paper made a systematic study on the *confidentiality* of IDAs and its connection with the adopted erasure codes [2]. Specially, this paper studied the confidentiality of IDAs from the view of practical security. This paper defined and discussed two levels of confidentiality: *weak confidentiality* (in the case where some parts of the original file can be reconstructed explicitly from fewer than the threshold number of pieces) and *strong confidentiality* (in the case where nothing of the original file can be reconstructed explicitly from fewer than the

threshold number of pieces). This paper showed which kind of IDAs have weak confidentiality and how an eavesdropper can reconstruct some segments of the original file explicitly from fewer than the threshold number of pieces in the case of weak confidentiality (see Theorem 3.1). It was noticed that for an IDA that adopts an arbitrary non-systematic erasure code, its confidentiality may fall into weak confidentiality. To achieve strong confidentiality, this paper explored a sufficient and feasible condition for an IDA (see Theorem 4.1). It was showed that Rabin's IDA [1] has strong confidentiality. At the same time, this paper presented an effective way to construct an IDA with strong confidentiality. Then, as an example, this paper constructed an IDA with strong confidentiality from a Reed-Solomon code [10], the computation complexity of which is comparable to or sometimes even lower than that of Rabin's IDA.

The key message we want to deliver through this paper is that *an arbitrary non-systematic erasure code is not enough for strong confidentiality in an IDA*. When referring to the confidentiality of an IDA in a practical application, we should keep this point in mind!

REFERENCES

- [1] M. O. Rabin, "Efficient dispersal of information for security, load balancing, and fault tolerance," *Journal of the ACM*, vol. 36, no. 2, pp. 335–348, Apr. 1989.
- [2] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. New York: North-Holland, 1977.
- [3] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, Oct. 1949.
- [4] G. R. Blakley, "Safeguarding cryptographic keys," in *Proceedings of the National Computer Conference*, New York, NY, Jun. 1979, pp. 313–317.
- [5] A. Shamir, "How to share a secret," *Communications of the ACM*, vol. 22, no. 11, pp. 612–613, Nov. 1979.
- [6] E. D. Karnin, J. W. Greene, and M. E. Hellman, "On secret sharing systems," *IEEE Transactions on Information Theory*, vol. 29, no. 1, pp. 35–41, Jan. 1983.
- [7] G. R. Blakley and C. Meadows, "Security of ramp schemes," in *Advances in Cryptology: Proceedings of CRYPTO '84*, ser. Lecture Notes in Computer Science, G. R. Blakley and D. Chaum, Eds. Springer-Verlag Berlin/Heidelberg, 1985, vol. 196, pp. 242–268.
- [8] A. D. Santis and B. Masucci, "Multiple ramp schemes," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1720–1728, Jul. 1999.
- [9] J. J. Wylie, M. W. Bigrigg, J. D. Strunk, G. R. Ganger, H. Kiliççote, and P. K. Khosla, "Survivable information storage systems," *Computer*, vol. 33, no. 8, pp. 61–68, Aug. 2000.
- [10] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, Jun. 1960.
- [11] J. Lacan and J. Fimes, "A construction of matrices with no singular square submatrices," in *Finite Fields and Applications*, ser. Lecture Notes in Computer Science, G. L. Mullen, A. Poli, and H. Stichtenoth, Eds. Springer-Verlag Berlin/Heidelberg, 2004, vol. 2948, pp. 145–147.
- [12] —, "Systematic MDS erasure codes based on Vandermonde matrices," *IEEE Communications Letters*, vol. 8, no. 9, pp. 570–572, Sep. 2004.



Mingqiang Li received a Ph.D. degree (with honor) in Computer Science from Tsinghua University in July, 2011. He also received a B.S. degree in Mathematics from the University of Electronic Science and Technology of China in July, 2006. He worked as a Staff Researcher in the IBM China Research Laboratory from July, 2011 to February, 2013. He is now a Postdoctoral Fellow in the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His current research interests include coding theory, storage systems, data security, data

compression, cloud infrastructure, distributed systems, wireless networking, and network economics.